Harvard Business Review

The Surprising Power of Online Experiments

by Ron Kohavi and Stefan Thomke (September-October 2017)

In 2012 a Microsoft employee working on Bing had an idea about changing the way the search engine displayed ad headlines. Developing it wouldn't require much effort—just a few days of an engineer's time—but it was one of hundreds of ideas proposed, and the program managers deemed it a low priority. So it languished for more than six months, until an engineer, who saw that the cost of writing the code for it would be small, launched a simple online controlled experiment—an A/B test—to assess its impact. Within hours the new headline variation was producing abnormally high revenue, triggering a "too good to be true" alert. Usually, such alerts signal a bug, but not in this case. An analysis showed that the change had increased revenue by an astonishing 12%—which on an annual basis would come to more than $100 million in the United States alone—without hurting key user-experience metrics. It was the best revenue-generating idea in Bing's history, but until the test its value was underappreciated.

Humbling! This example illustrates how difficult it can be to assess the potential of new ideas. Just as important, it demonstrates the benefit of having a capability for running many tests cheaply and concurrently—something more businesses are starting to recognize.

Today, Microsoft and several other leading companies—including Amazon, Booking.com, Facebook, and Google—each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users. Start-ups and companies without digital roots, such as Walmart, Hertz, and Singapore Airlines, also run them regularly, though on a smaller scale. These organizations have discovered that an "experiment with everything" approach has surprisingly large payoffs. It has helped Bing, for instance, identify dozens of revenue-related changes to make each month—improvements that have collectively increased revenue per search by 10% to 25% each year. These enhancements, along with hundreds of other changes per month that increase user satisfaction, are the major reason that Bing is profitable and that its share of U.S. searches conducted on personal computers has risen to 23%, up from 8% in 2009, the year it was launched.

At a time when the web is vital to almost all businesses, rigorous online experiments should be standard operating procedure. If a company develops the software infrastructure and organizational skills to conduct them, it will be able to assess not only ideas for websites but also potential business models, strategies, products, services, and marketing campaigns—all relatively inexpensively. Controlled experiments can transform decision making into a scientific, evidence-driven process—rather than an intuitive reaction. Without them, many breakthroughs might never happen, and many bad ideas would be implemented, only to fail, wasting resources.

Yet we have found that too many organizations, including some major digital enterprises, are haphazard in their experimentation approach, don't know how to run rigorous scientific tests, or conduct way too few of them.

Together we've spent more than 35 years studying and practicing experiments and advising companies in a wide range of industries about them. In these pages we'll share the lessons we've gleaned about how to design and execute them, ensure their integrity, interpret their results, and address the challenges they're likely to pose. Though we'll focus on the simplest kind of controlled experiment, the A/B test, our findings and suggestions apply to more-complex experimental designs as well.

Appreciate the Value of A/B Tests

In an A/B test the experimenter sets up two experiences: "A," the control, is usually the current system and considered the "champion," and "B," the treatment, is a modification that attempts to improve something—the "challenger." Users are randomly assigned to the experiences, and key metrics are computed and compared. (Univariable A/B/C tests and A/B/C/D tests and multivariable tests, in contrast, assess more than one treatment or modifications of different variables at the same time.) Online, the modification could be a new feature, a change to the user interface (such as a new layout), a back-end change (such as an improvement to an algorithm that, say, recommends books at Amazon), or a different business model (such as an offer of free shipping). Whatever aspect of operations companies care most about—be it sales, repeat usage, click-through rates, or time users spend on a site—they can use online A/B tests to learn how to optimize it.

Any company that has at least a few thousand daily active users can conduct these tests. The ability to access large customer samples, to automatically collect huge amounts of data about user interactions on websites and apps, and to run concurrent experiments gives companies an unprecedented opportunity to evaluate many ideas quickly, with great precision, and at a negligible cost per incremental experiment. That allows organizations to iterate rapidly, fail fast, and pivot.

Recognizing these virtues, some leading tech companies have dedicated entire groups to building, managing, and improving an experimentation infrastructure that can be employed by many product teams. Such a capability can be an important competitive advantage—provided you know how to use it. Here's what managers need to understand:

Tiny changes can have a big impact.

People commonly assume that the greater an investment they make, the larger an impact they'll see. But things rarely work that way online, where success is more about getting many small changes right. Though the business world glorifies big, disruptive ideas, in reality most progress is achieved by implementing hundreds or thousands of minor improvements.

Putting credit card offers on the shopping cart page boosted profits by millions.

Consider the following example, again from Microsoft. (While most of the examples in this article come from Microsoft, where Ron heads experimentation, they illustrate lessons drawn from many companies.) In 2008 an employee in the United Kingdom made a seemingly minor suggestion: Have a new tab (or a new window in older browsers) automatically open whenever a user clicks on the Hotmail link on the MSN home page, instead of opening Hotmail in the same tab. A test was run with about 900,000 UK users, and the results were highly encouraging: The engagement of users who opened Hotmail increased by an impressive 8.9%, as measured by the number of clicks they made on the MSN home page. (Most changes to engagement have an effect smaller than 1%.) However, the idea was controversial because few sites at the time were opening links in new tabs, so the change was released only in the UK.

In June 2010 the experiment was replicated with 2.7 million users in the United States, producing similar results, so the change was rolled out worldwide. Then, to see what effect the idea might have elsewhere, Microsoft explored the possibility of having people who initiated a search on MSN open the results in a new tab. In an experiment with more than 12 million users in the United States, clicks per user increased by 5%. Opening links in new tabs is one of the best ways to increase user engagement that Microsoft has ever introduced, and all it required was changing a few lines of code. Today many websites, including Facebook.com and Twitter.com, use this technique.

Microsoft's experience is hardly unique. Amazon's experiments, for instance, revealed that moving credit card offers from its home page to the shopping cart page boosted profits by tens of millions of dollars annually. Clearly, small investments can yield big returns. Large investments, however, may have little or no payoff. Integrating Bing with social media—so that content from Facebook and Twitter opened on a third pane on the search results page—cost Microsoft more than $25 million to develop and produced negligible increases in engagement and revenue.

Experiments can guide investment decisions.

Online tests can help managers figure out how much investment in a potential improvement is optimal. This was a decision Microsoft faced when it was looking at reducing the time it took Bing to display search results. Of course, faster is better, but could the value of an improvement be quantified? Should there be three, 10, or perhaps 50 people working on that performance enhancement? To answer those questions, the company conducted a series of A/B tests in which artificial delays were added to study the effects of minute differences in loading speed. The data showed that every 100-millisecond difference in performance had a 0.6% impact on revenue. With Bing's yearly revenue surpassing $3 billion, a 100-millisecond speedup is worth $18 million in annual incremental revenue—enough to fund a sizable team.

The test results also helped Bing make important trade-offs, specifically about features that might improve the relevance of search results but slow the software's response time. Bing wanted to avoid a situation in which many small features cumulatively led to a significant degradation in performance. So the release of individual features that slowed the response by more than a few milliseconds was delayed until the team improved either their performance or the performance of another component.

Build a Large-Scale Capability

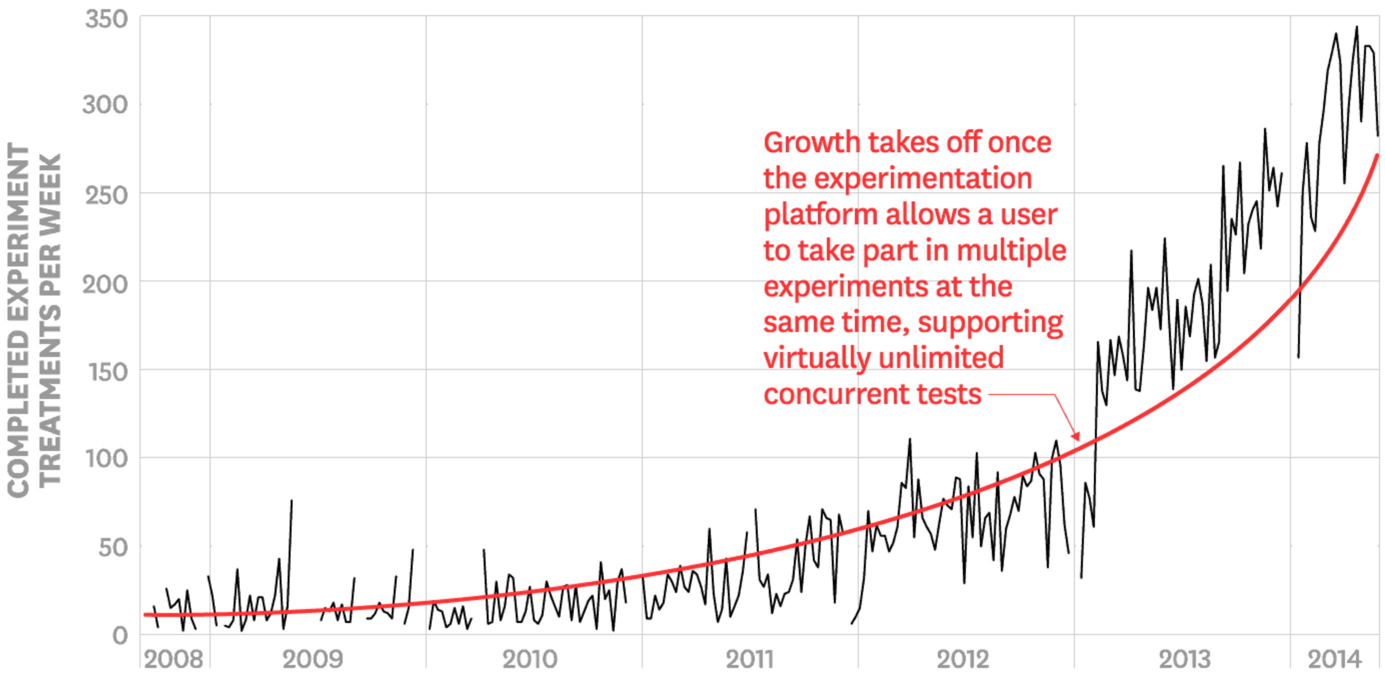
More than a century ago, the department store owner John Wanamaker reportedly coined the marketing adage "Half the money I spend on advertising is wasted; the trouble is that I don't know which half." We've found something similar to be true of new ideas: The vast majority of them fail in experiments, and even experts often misjudge which ones will pay off. At Google and Bing, only about 10% to 20% of experiments generate positive results. At Microsoft as a whole, one-third prove effective, one-third have neutral results, and one-third have negative results. All this goes to show that companies need to kiss a lot of frogs (that is, perform a massive number of experiments) to find a prince.

Any figure that looks interesting or different is usually wrong.

It's key to experiment with everything to make sure that changes neither are degrading nor have unexpected effects. At Bing about 80% of proposed changes are first run as controlled experiments. (Some low-risk bug fixes and machine-level changes like operating system upgrades are excluded.)

Scientifically testing nearly every proposed idea requires an infrastructure: instrumentation (to record such things as clicks, mouse hovers, and event times), data pipelines, and data scientists. Several third-party tools and services make it easy to try experiments, but if you want to scale things up, you must tightly integrate the capability into your processes. That will drive down the cost of each experiment and increase its reliability. On the other hand, a lack of infrastructure will keep the marginal costs of testing high and could make senior managers reluctant to call for more experimentation.

The Growth of Experimentation at Bing



Growth takes off once the experimentation platform allows a user to take part in multiple experiments at the same time, supporting virtually unlimited concurrent tests

Microsoft provides a good example of a substantial testing infrastructure—though a smaller enterprise or one whose business is not as dependent on the experimentation could make do with less, of course. Microsoft's Analysis & Experimentation team consists of more than 80 people who on any given day help run hundreds of online controlled experiments on various products, including Bing, Cortana, Exchange, MSN, Office, Skype, Windows, and Xbox. Each experiment exposes hundreds of thousands—and sometimes even tens of millions—of users to a new feature or change. The team runs rigorous statistical analyses on all these tests, automatically generating scorecards that check hundreds to thousands of metrics and flag significant changes.

A company's experimentation personnel can be organized in three ways:

Centralized model.

In this approach a team of data scientists serve the entire company. The advantage is that they can focus on long-term projects, such as building better experimentation tools and developing more-advanced statistical algorithms. One major drawback is that the business units using the group may have different priorities, which could lead to conflicts over the allocation of resources and costs. Another con is that data scientists may feel like outsiders when dealing with the businesses and thus be less attuned to the units' goals and domain knowledge, which could make it harder for them to connect the dots and share relevant insights. Moreover, the data scientists may lack the clout to persuade senior management to invest in building the necessary tools or to get corporate and business unit managers to trust the experiments' results.

Decentralized model.

Another approach is distributing data scientists throughout the different business units. The benefit of this model is that the data scientists can become experts in each business domain. The main disadvantage is the lack of a clear career path for these professionals, who also may not receive peer feedback and mentoring that help them develop. And experiments in individual units may not have the critical mass to justify building the required tools.

Center-of-excellence model.

A third option is to have some data scientists in a centralized function and others within the different business units.

(Microsoft uses this approach.) A center of excellence focuses mostly on the design, execution, and analysis of controlled experiments. It significantly lowers the time and resources those tasks require by building a companywide experimentation platform and related tools. It can also spread best testing practices throughout the organization by hosting classes, labs, and conferences. The main disadvantages are a lack of clarity about what the center of excellence owns and what the product teams own, who should pay for hiring more data scientists when various units increase their experiments, and who is responsible for investments in alerts and checks that indicate results aren't trustworthy.

There is no right or wrong model. Small companies typically start with the centralized model or use a third-party tool and then, after they've grown, switch to one of the other models. In companies with multiple businesses, managers who consider testing a priority may not want to wait until corporate leaders develop a coordinated organizational approach; in those cases, a decentralized model might make sense, at least in the beginning. And if online experimentation is a corporate priority, a company may want to build expertise and develop standards in a central unit before rolling them out in the business units.

Address the Definition of Success

Every business group must define a suitable (usually composite) evaluation metric for experiments that aligns with its strategic goals. That might sound simple, but determining which short-term metrics are the best predictors of long-term outcomes is difficult. Many companies get it wrong. Getting it right—coming up with an *overall evaluation criterion* (OEC)—takes thoughtful consideration and often extensive internal debate. It requires close cooperation between senior executives who understand the strategy and data analysts who understand metrics and trade-offs. And it's not a onetime exercise: We recommend that the OEC be adjusted annually.

Arriving at an OEC isn't straightforward, as Bing's experience shows. Its key long-term goals are increasing its share of search-engine queries and its ad revenue. Interestingly, decreasing the relevance of search results will cause users to issue more queries (thus increasing query share) and click more on ads (thus increasing revenue). Obviously, such gains would only be short-lived, because people would eventually switch to other search engines. So which short-term metrics do predict long-term improvements to query share and revenue? In their discussion of the OEC, Bing's executives and data analysts decided that they wanted to *minimize* the number of user queries for each task or session and *maximize* the number of tasks or sessions that users conducted.

It's also important to break down the components of an OEC and track them, since they typically provide insights into why an idea was successful. For example, if number of clicks is integral to the OEC, it's critical to measure which parts of a page were clicked on. Looking at different metrics is crucial because it helps teams discover whether an experiment has an unanticipated impact on another area. For example, a team making a change to the related search queries shown (a search on, say, "Harry Potter," will show queries about Harry Potter books, Harry Potter movies, the casts of those movies, and so on) may not realize that it's altering the distribution of queries (by increasing searches for the related queries), which could affect revenue positively or negatively.

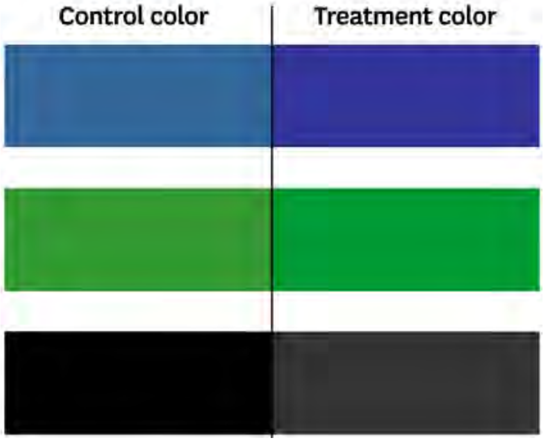
Over time the process of building and adjusting the OEC and understanding causes and effects becomes easier. By running experiments, debugging the results (which we will discuss in a little bit), and interpreting them, companies will not only gain valuable experience with what metrics work best for certain types of tests but also develop new metrics. Over the years, Bing has created more than 6,000 metrics experimenters can use, which are grouped into templates by the area the tests involve (web search, image search, video search, changes to ads, and so on).

Beware of Low-Quality Data

It doesn't matter how good your evaluation criteria are if people don't trust experiments' results. Getting numbers is easy; getting numbers you can trust is hard! You need to allocate time and resources to validating the experimentation system and setting up automated checks and safeguards. One method is to run rigorous A/A tests—that is, test something against itself to ensure that about 95% of the time the system correctly identifies no statistically significant difference. This simple approach has helped Microsoft identify hundreds of invalid experiments and improper applications of formulas (such as using a formula that assumes all measurements are independent when they are not).

Small Changes with a Huge Impact

Bing's experiments showed that slightly darker blues and greens in titles and a slightly lighter black in captions improved the users' experience. When rolled out to all users, the color changes boosted revenue by more than $10 million annually.



We've learned that the best data scientists are skeptics and follow Twyman's law: Any figure that looks interesting or different is usually wrong. Surprising results should be replicated—both to make sure they're valid and to quell people's doubts. In 2013, for example, Bing ran a set of experiments with the colors of various text that appeared on its search results page, including titles, links, and captions. Though the color changes were subtle, the results were unexpectedly positive: They showed that users who saw slightly darker blues and greens in titles and a slightly lighter black in captions were successful in their searches a larger percentage of the time and that those who found what they wanted did so in significantly less time.

Since the color differences are barely perceptible, the results were understandably viewed with skepticism by multiple disciplines, including the design experts. (For years, Microsoft, like many other companies, had relied on expert designers—rather than the behavior of actual users—to define corporate style guides and colors.) So the experiment was rerun with a much larger sample of 32 million users, and the results were similar. Analysis indicated that when rolled out to all users, the color changes would increase revenue by more than $10 million annually.

CONCLUSION

If you want results to be trustworthy, you must ensure that high-quality data is used. Outliers may need to be excluded, collection errors identified, and so on. In the online world this issue is especially important, for several reasons. Take internet bots. At Bing more than 50% of requests come from bots. That data can skew results or add "noise," which makes it harder to detect statistical significance. Another problem is the prevalence of outlier data points. Amazon, for instance, discovered that certain individual users made massive book orders that could skew an entire A/B test; it turned out they were library accounts.

Managers should also beware when some segments experience much larger or smaller effects than others do (a phenomenon statisticians call "heterogeneous treatment effects"). In certain cases a single good or bad segment can skew the average enough to invalidate the overall results. This happened in one Microsoft experiment in which one segment, Internet Explorer 7 users, couldn't click on the results of Bing searches because of a JavaScript bug, and the overall results, which were otherwise positive, turned negative. An experimentation platform should detect such unusual segments; if it doesn't, experimenters looking at an average effect may dismiss a good idea as a bad one.

Results may also be biased if companies reuse control and treatment populations from one experiment to another. That practice leads to "carryover effects," in which people's experience in an experiment alters their future behavior. To avoid this phenomenon, companies should "shuffle" users between experiments.

Another common check Microsoft's experimentation platform performs is validating that the percentages of users in the control and treatment groups in the actual experiment match the experimental design. When these differ, there is a "sample ratio mismatch," which often voids the results. For example, a ratio of 50.2/49.8 (821,588 versus 815,482 users) diverges enough from an expected 50/50 ratio that the probability that it happened by chance is less than one in 500,000. Such mismatches occur regularly (usually weekly), and teams need to be diligent in understanding why and resolving them.

Avoid Assumptions About Causality

Because of the hype over big data, some executives mistakenly believe that causality isn't important. In their minds all they need to do is establish correlation, and causality can be inferred. Wrong!

The following two examples illustrate why—and also highlight the shortcomings of experiments that lack control groups. The first concerns two teams that conducted separate observational studies of two advanced features for Microsoft Office. Each concluded that the new feature it was assessing reduced attrition. In fact, almost any advanced feature will show such a correlation, because people who will try an advanced feature tend to be heavy users, and heavy users tend to have lower attrition. So while a new advanced feature might be correlated with lower attrition, it doesn't necessarily cause it. Office users who get error messages also have lower attrition, because they too tend to be heavy users. But does that mean that showing users more error messages will reduce attrition? Hardly.

The second example concerns a study Yahoo did to assess whether display ads for a brand, shown on Yahoo sites, could increase searches for the brand name or related keywords. The observational part of the study estimated that the ads increased the number of searches by 871% to 1,198%. But when Yahoo ran a controlled experiment, the increase was only 5.4%. If not for the control, the company might have concluded that the ads had a huge impact and wouldn't have realized that the increase in searches was due to other variables that changed during the observation period.

Some executives believe that all they need to do is establish correlation. Wrong!

Clearly, observational studies cannot establish causality. This is well known in medicine, which is why the U.S. Food and Drug Administration mandates that companies conduct randomized clinical trials to prove that their drugs are safe and effective.

Including too many variables in tests also makes it hard to learn about causality. With such tests it's difficult to disentangle results and interpret them. Ideally, an experiment should be simple enough that cause-and-effect relationships can be easily understood. Another downside of complex designs is that they make experiments much more vulnerable to bugs. If a new feature has a 10% chance of triggering an egregious problem that requires aborting its test, then the probability that a change that involves seven new features will have a fatal bug is more than 50%.

What if you can determine that one thing causes another, but you don't know why? Should you try to understand the causal mechanism? The short answer is yes.

Between 1500 and 1800, about 2 million sailors died of scurvy. Today we know that scurvy is caused by a lack of vitamin C in the diet, which sailors experienced because they didn't have adequate supplies of fruit on long voyages. In 1747, Dr. James Lind, a surgeon in the Royal Navy, decided to do an experiment to test six possible cures. On one voyage he gave some sailors oranges and lemons, and others alternative remedies like vinegar. The experiment showed that citrus fruits could prevent scurvy, though no one knew why. Lind mistakenly believed that the acidity of the fruit was the cure and tried to create a less-perishable remedy by heating the citrus juice into a concentrate, which destroyed the vitamin C. It wasn't until 50 years later, when unheated lemon juice was added to sailors' daily rations, that the Royal Navy finally eliminated scurvy among its crews. Presumably, the cure could have come much earlier and saved many lives if Lind had run a controlled experiment with heated and unheated lemon juice.

That said, we should point out that you don't always have to know the "why" or the "how" to benefit from knowledge of the "what." This is particularly true when it comes to the behavior of users, whose motivations can be difficult to determine. At Bing some of the biggest breakthroughs were made without an underlying theory. For example, even though Bing was able to improve the user experience with those subtle changes in the colors of the type, there are no well-established theories about color that could help it understand why. Here the evidence took the place of theory.

The online world is often viewed as turbulent and full of peril, but controlled experiments can help us navigate it. They can point us in the right direction when answers aren't obvious or people have conflicting opinions or are uncertain about the value of an idea.

Several years ago, Bing was debating whether to make ads larger so that advertisers could include links to specific landing pages in them. (For example, a loan company might provide links like "compare rates" and "about the company" instead of just one to a home page.) A downside was that larger ads obviously would take up more screen real estate, which is known to increase user dissatisfaction and churn. The people considering the idea were split. So the Bing team experimented with increasing the ads' size while keeping the overall screen space allotted for ads constant, which meant showing fewer of them. The upshot was that showing fewer but larger ads led to a big improvement: Revenue increased by more than $50 million annually without hurting the key aspects of the user experience.

If you really want to understand the value of an experiment, look at the difference between its expected outcome and its actual result. If you thought something was going to happen and it happened, then you haven't learned much. If you thought something was going to happen and it didn't, then you've learned something important. And if you thought something minor was going to happen, and the results are a major surprise and lead to a breakthrough, you've learned something highly valuable.

By combining the power of software with the scientific rigor of controlled experiments, your company can create a learning lab. The returns you reap—in cost savings, new revenue, and improved user experience—can be huge. If you want to gain a competitive advantage, your firm should build an experimentation capability and master the science of conducting online tests.

A version of this article appeared in the September–October 2017 issue (pp.74–82) of *Harvard Business Review*.

Ron Kohavi is a distinguished engineer and the general manager of the Analysis & Experimentation team at Microsoft. Previously, he was the director of data mining and personalization at Amazon, where he was responsible for Weblab, Amazon's experimentation system.

Stefan Thomke is the William Barclay Harding Professor of Business Administration at Harvard Business School. He is a leading authority on the management of business experimentation and innovation and has worked with many global companies on product, process, and technology development.